Computational and Structural Biotechnology Journal 27 (2025) 744–752

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Research Article

SIN-Seg: A joint spatial-spectral information fusion model for medical image segmentation

Siyuan Dai [a], Kai Ye [a], Charlie Zhan [a], Haoteng Tang [b,*], Liang Zhan [a,*]

[a] Department of Electrical and Computer Engineering, University of Pittsburgh, Pittsburgh, 15213, PA, USA
[b] Department of Computer Science, University of Texas Rio Grande Valley, Edinburg, 78582, TX, USA

ARTICLE INFO

Keywords:
Spectral information
Feature alignment
Medical image segmentation
Contrastive learning

ABSTRACT

In recent years, the application of deep convolutional neural networks (DCNNs) to medical image segmentation has shown significant promise in computer-aided detection and diagnosis (CAD). Leveraging features from different spaces (i.e. Euclidean, non-Euclidean, and spectrum spaces) and multi-modalities of data have the potential to improve the information available to the CAD system, enhancing both effectiveness and efficiency. However, directly acquiring data from different spaces across multi-modalities is often prohibitively expensive and time-consuming. Consequently, most current medical image segmentation techniques are confined to the spatial domain, which is limited to utilizing scanned images from MRI, CT, PET, etc. Here, we introduce an innovative Joint Spatial-Spectral Information Fusion method which requires no additional data collection for CAD. We translate existing single-modality data into a new domain to extract features from an alternative space. Specifically, we apply Discrete Cosine Transformation (DCT) to enter the spectrum domain, thereby accessing supplementary feature information from an alternate space. Recognizing that information from different spaces typically necessitates complex alignment modules, we introduce a contrastive loss function for achieving feature alignment before synchronizing information across different feature spaces. Our empirical results illustrate the greater effectiveness of our model in harnessing additional information from the spectrum-based space and affirm its superior performance against influential state-of-the-art segmentation baselines. The code is available at https://github.com/Auroradsy/SIN-Seg.

1. Introduction

Medical image segmentation is a critical component in the fields of biomedical science research and clinical diagnosis. Its goal is to delineate regions of interest (ROIs) that possess significant diagnostic and therapeutic value for treating physicians and radiologists. The advent of computer-aided detection/diagnosis (CAD) systems has facilitated a unified platform for analyzing vast amounts of medical-specific imaging data. (i.e. MRI, CT, Microscopy, PET, etc) Within this framework, deep neural networks (DNNs) based models have showcased their value, offering precise segmentation outcomes and reducing the time burden traditionally associated with manual analysis.

Despite the impressive achievements of DNNs [1–6], intrinsic challenges remain to the methodologies currently in medical image segmentation. Medical images, acquired through various specialized devices, are designed to accentuate particular features or abnormalities, often requiring extra interpretative expertise of radiologists to achieve precise diagnosis. A typical CAD system that operates on images from a single type of information, without integrating such expert insight, risks overlooking critical information. Multi-modal learning in medical image analysis [7] can harness the strengths of diverse imaging modalities—such as MRI, CT, and PET to improve diagnostic accuracy over single-modality data. However, collecting multi-modal data for a single subject using different imaging devices is time-consuming and expensive in practical situations. Even though MRI devices can produce images in multiple modalities by capturing different sequence scans in a single session, potentially enhancing diagnostic effectiveness [8], such scanning processes require skilled radiologists or technicians and involve setting up various MRI contrast media. This is not only time-intensive but also incurs significant costs. Moreover, multiple imaging modalities require patients to be exposed to radiation from MRI devices. Typical MRI imaging is diagnosis-oriented, and regular MRI images aim at specific requirements and are captured under particular sequences.

To address this issue, we propose a novel spectrum space-based Joint Spatial-Spectral Information Fusion model (SIN). Prior researchers [9,10] have illustrated the benefits of spectrum domain learning, par-

* Corresponding authors.

https://doi.org/10.1016/j.csbj.2025.02.024
Received 15 October 2024; Received in revised form 21 February 2025; Accepted 21 February 2025

Available online 24 February 2025

ticularly in edge detection—a critical element of segmentation tasks. These studies have established the validity and significance of spectral information from the frequency domain in augmenting image contrast and delineating abnormalities and pathological regions. Spectral information is particularly pivotal in MRI, CT, and microscopy, such as frequency sequence-related imaging, where it reveals highly distinctive features of the same segmentation target under varied spectral-related settings during data acquisition [11]. Specifically, high-frequency features may be overlooked in the spatial domain, whereas these features are more readily extracted in the spectral domain [10].

Our SIN model innovatively harnesses both spectral and spatial domain information, synthesizing features from these two spaces. It comprises two primary components: an offline discrete cosine transform (DCT) module and an online trainable feature alignment module, both of which are embeddable and compatible with every encoder-decoder-based segmentation architecture and maintain the end-to-end attributes. The DCT transformation is color-sensitive and microscopy images are captured under *RGB* color space, unlike other medical images which are in the *gray-scale* color space. For those three channel-based microscopy images, we implement a space transformation from the *RGB* color space to the *YCbCr* color space, leveraging the fact that the feature is more sensitive to changes in brightness than color changes, resulting in a more efficient form of further image processing.

Furthermore, more feature modalities [12] and larger parameter-based models [13] could enhance the performance of models. However, aligning features from disparate domains or modalities, each rooted in different spaces presents a significant challenge, often necessitating complex modules for integration [14,15]. In this study, our proposed model does not target solving multimodal feature fusion challenges so we designed our model with fewer parameters to be used for feature alignment. To overcome this, we introduced a contrastive learning strategy to align the features inspired by CLIP [16].

In summary, our main contributions to this paper can be shown as follows:

- We propose a novel dual information extraction framework for fusing the information both from the spectral and the spatial feature space.
- We introduce a low-dimension flattened strategy for the information from different feature spaces, combined with a simple contrastive loss for feature alignment which does not need extra parameters.
- We verify our proposed model on multiple datasets from different medical imaging devices, involving a brain tumor segmentation dataset [17] and a heart segmentation dataset [18] both from MRI devices, a liver segmentation dataset [19] captured under CT devices, and a cell segmentation dataset [20] from different microscopy imaging methods. We compared our proposed framework with 8 influential single UNet and Transformer-based baselines to show the superiority. We also highlight its effectiveness and potential for advancing medical image analysis under other comprehensive analysis experiments.

2. Related work

In this section, we briefly review the previous works in three different aspects highly related to our works. First, we introduce the improvement of the backbone model. Then, how spectral information shows its significance and potential in the computer vision tasks. Finally, some previous works about how to combine and take advantage of different feature spaces with alignment strategies are illustrated.

2.1. Foundation models for medical image segmentation

Semantic segmentation is always a crucial task for the computer vision domain. FCN [1]] is the first research that introduced Convolutional Neural Network (CNN) for segmentation. Then, UNet [2] took advantage of the encoder-decoder-like architecture, initially introduced for biomedical image segmentation, and revolutionized the field of medical image analysis. Its unique design, characterized by a symmetric downsample and upsample path, combined with a skip connection allows for precise localization and context capture, resulting in highly effective performance in the medical image domain. Based on such a powerful foundation model, researchers proposed numerous advanced segmentation frameworks. Zhang, et al. [21] introduced a hard attention mechanism to UNet and utilized the superiority of ResNet, proposing Res-UNet. Then, Att-UNet [22] was designed with a gate module for soft attention calculation to enhance the performance of the original UNet. Zhou, et al. [23] and Valanarasu, et al. [24] also proposed novel models based on UNe. Then, TransUNet [25] introduced a vision transformer (ViT) [26,27] after a down-sample which could combine the spatial semantic and the local semantic in consideration in the hidden space. Additionally, researchers [28,29] modified the ViT structure and showed the significance of a new Transformer-based foundation model in medical image segmentation.

2.2. Spectral information

Conventional computer vision algorithms mainly consider the image analysis in the spatial space, i.e. the *RGB* or *Gray-Scale* images which are easily recognized by human eyes. However, the information in such space could obscure lots of detailed features. Some research works [30,31] have found that when processing a visual scene, animals have more wavebands than humans because of their unique ability to spot the features in the spectral domain. Therefore, the significant semantic information is easier to extract in such a feature space, and utilizing the transformed features in the spectral space is a sufficient method to compress the images, [32,33,10] and also design lighter networks [34–36] themselves. It is also natural to take advantage of the spectral information for designing the attention pipelines. Qin, et al. [37] found that many works have used global average pooling (GAP) as an unquestionable preprocessing method for designing channel attention mechanisms. A potential problem is that different channels may have the same mean value, while their corresponding semantic information may be completely different, which creates the problem of insufficient attention information. They proved that GAP is a special case of DCT, which is equivalent to the lowest frequency component of DCT and is generalized to the frequency domain, proposing a multi-spectral channel attention framework. Meanwhile, FRCU-Net [38] also incorporates a Laplacian transformation-based method to compute attention and enhance feature calibration, resulting in improved performance compared to the vanilla UNet [2]. FSDR [39] introduced a novel attention pipeline based on the spectral space for forcing the network to learn more intrinsic semantic features and achieve a more generalizable model. Additionally, spectral space could be highly beneficial on some low-level vision scene tasks [40,41], implementing the DCT or Wavelet transformation. XNet [42] effectively integrates high-frequency and low-frequency features using Wavelet transformation, demonstrating superior performance in medical image segmentation under both fully and semi-supervised learning paradigm.

2.3. Feature alignment

Humans perceive the world through different organs, and more information could train more powerful neural networks. Nevertheless, the information in the different spaces always obstacles each other before aligning into the same feature space. For naive feature fuse, e.g. *Concatenation* is always considered, and it is widely used in residual block, skip-connection, and related fusion situations. Under such an operation, multiple feature maps are spliced together in the depth dimension to obtain a richer representation of features. For example, in encoders and decoders, low-level features and high-level features are spliced, which

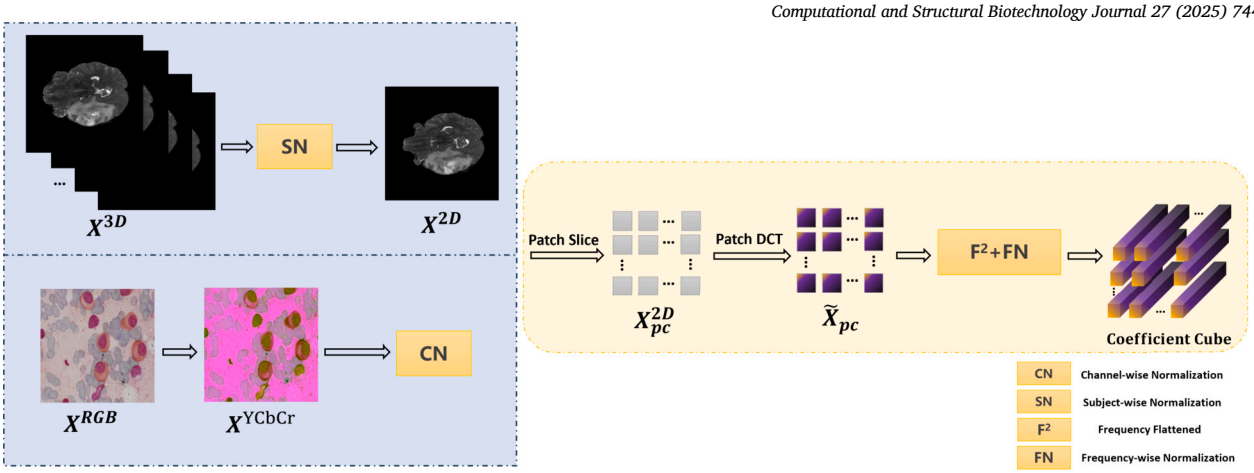


Fig. 1. An overview of the off-line DCT transformation module in the SIN model for *RGB* space. First, an *RGB* image is converted to the *YCbCr* domain. Then the *YCbCr* image is divided into small image patches with a channel-wise normalization (CN). Next, a DCT transformation is implemented on image patches. Finally, the coefficient cube for the whole image is generated from frequency-based flattened (F^2) and frequency-wise normalization (FN) operations.

improves the decoder's perceptual ability. However, concatenation-like fusion methods are not trainable, which pressures other learnable blocks to handle the feature fusion. Multimodal learning fields often meet such problems [43,44]. Autonomous driving [45] is a typical computer vision task that requires multi-modal features. In the ordinary road environment with traffic lights, traffic cones, etc., relying on information from a single modality is insufficient while fusing features from RGB, LiDAR, Text, et al. Tan, et al.. [46,14] attempted to introduce multimodal-learning in the medical image segmentation task by using different medical imaging devices on an organ and utilizing the information from different modalities. Unfortunately, the complex modules used to handle the feature fusion challenges consumed too much computational memory. Contrastive learning [47,48] is a simple and efficient method to align and merge data or feature maps from different feature spaces. Inspired by multimodal contrastive learning, we introduce a simple contrastive learning strategy to achieve feature alignment without extra parameters.

3. Methodology

In this section, we present our proposed SIN model which introduces the spectral information and integrates it with spatial information for segmentation tasks. We first propose a novel off-line DCT transformation module in Fig. 1 to convert the image from the spatial space to the spectrum space. We then introduce a trainable alignment module with a simple contrastive loss function to align the features yielded from the spectral space and the spatial space as well. Finally, we illustrate the whole segmentation framework (named SIN-Seg) in Fig. 2 with our proposed SIN model and the loss functions for brain tumor segmentation tasks.

3.1. Off-line DCT transformation

3.1.1. RGB images pre-processing

Microscopy Imaging always generates into the *RGB* space, which is not suitable for conducting DCT transformation directly on *RGB* images (denoted as X^{RGB}). Instead, we first transform them to the *YCbCr* space as *YCbCr* images (denoted as X^{YCbCr}). This conversion is crucial for two main reasons: Human Visual Sensitivity: YCbCr separates an image into luminance (Y) and chrominance (Cb and Cr). Since human vision is more sensitive to luminance than chrominance, this separation allows for more effective compression. The luminance channel can be preserved with higher fidelity, while the chrominance channels can be compressed more, reducing file size without noticeably impacting image quality. Compression Efficiency: The DCT is more effective in the YCbCr space for compression purposes. It allows for significant data reduction in the chrominance components, which is less perceptible to the human eye while maintaining the crucial details in the luminance component. After such pre-processing, the color information of luminance and chrominance is separated into three channels including *Y* (i.e., luma or brightness), *Cb* (i.e., blue-difference chroma), *Cr* (i.e., red-difference chroma). The *YCbCr* transformation leverages the fact that the human visual system is more sensitive to changes in brightness than color changes, resulting in more efficient image processing. To implement the *YCbCr* transformation, we first normalize image *RGB* values to the range of $[0, 1]$ with their own min-max values subjects by subjects because of the difference intensity scale for different capturing institutions, and then convert the normalized *RGB* values to the *YCbCr* color space as follows:

$$X^{YCbCr} = \begin{cases} Y & = 0.299R + 0.587G + 0.114B \\ Cb & = -0.169R - 0.331G + 0.500B + 0.5 \\ Cr & = 0.500R - 0.419G - 0.081B + 0.5 \end{cases} \quad (1)$$

where *R*, *G*, and *B* represent the intensity values in the three channels (i.e., red, green, blue) of *RGB* images, respectively, while *Y*, *Cb*, and *Cr* represent the intensity values in the three channels of *YCbCr* images. The whole pipeline under *RGB* space for DCT transformation is shown as Fig. 1. So that we could get the *YCbCr* images (i.e., $X_{pc}^{YCbCr} \in \mathcal{R}^{H\times W\times C}$, where H and W denote image size, and C denotes the three channels of such color space) are generated, and the feature map in every channel will be implemented DCT transformation channel by channel.

3.1.2. Gray-scale images pre-processing

MRI and CT images are inherently grayscale, making them unaffected by variations in luminance and chromaticity. Consequently, there is no need to convert them to the YCbCr color space, and their original intensity representation remains suitable for analysis. Furthermore, since these images are acquired as 3D volumetric data rather than conventional RGB images, their intensity values do not necessarily conform to a standardized scale, such as the typical $[0, 255]$ range used in digital imaging. Since these types of datasets are collected by different institutions from different patients, we normalize them one patient by one patient with the min-max value of themselves to do subject-wise min-max normalization (SN), mapping them to the same intensity scale in the spatial domain. After normalization, we slice all the 3D MRI volumes into 2D images (i.e., $X_{pc}^{2D} \in \mathcal{R}^{H\times W}$, where H and W denote image size), and we conduct DCT transformation to convert them into the spectrum domain for another modality.

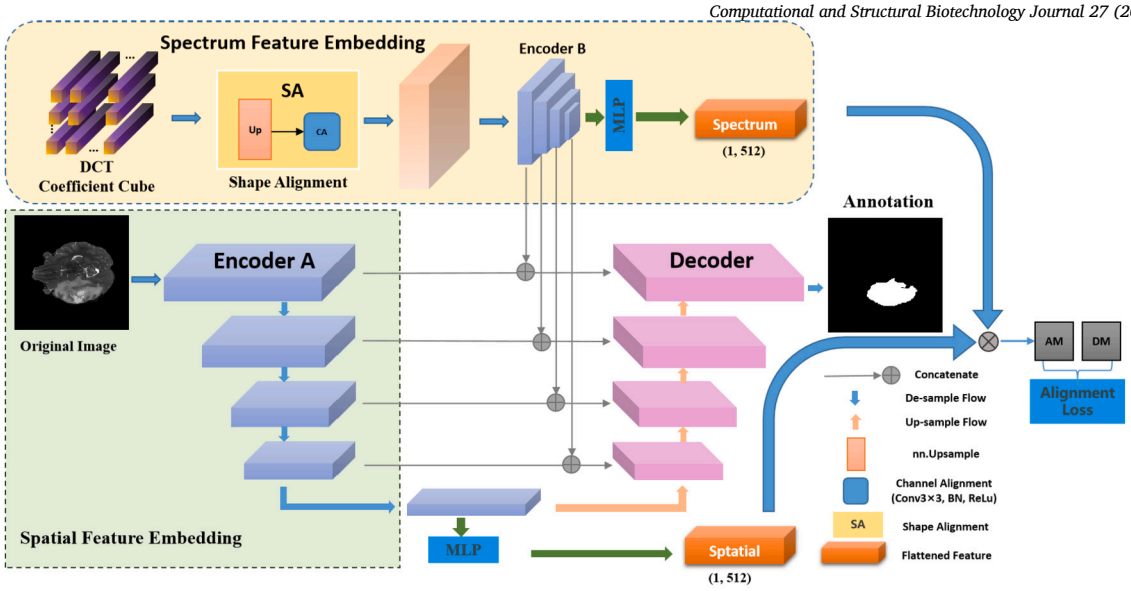


Fig. 2. Diagram of the proposed SIN-Seg framework, including two U-Net encoders for the original image in the spatial space and the DCT coefficient cube embedding in the spectrum space, respectively. The coefficient cube is first up-sampled and channel adjusted via the shape-alignment process, to make the input shape aligned to the feature in the spatial space. The features from both encoders are synchronized scale-by-scale. The fused features are then fed forward to the *U-Net* decoder to generate the final predicted segmentation masks. A feature alignment is also implemented on the flattened frequency, and spatial latent features are implemented with the alignment loss.

3.1.3. DCT transformation in patches

Particularly, the DCT transformation is conducted on the 8×8 patches of *2D* images, to extract more fine-grained features in the spectrum domain. The DCT transformation (i.e., $\tilde{X}_{pc} \in \mathcal{R}^{1\times 8\times 8}$) on every *2D* image patch is computed as follows:

$$\tilde{X}_{pc}(i,\ j) = \frac{2}{\sqrt{(N_1,\ N_2)}} \sum_{n_1=0}^{N_1-1} \sum_{n_2=0}^{N_2-1} X^{2D}(n_1,\ n_2) \cdot a_{n1} a_{n2} cos\left[n_1 \frac{2\pi}{/*-}(n_1+\frac{1}{2})\right] cos\left[n_2 \frac{2\pi}{N_2}(n_2+\frac{1}{2})\right] \quad (2)$$

$$s.t.\ a_{n1},\ a_{n2} = \begin{cases} \frac{1}{\sqrt{2}}, & k=0 \\ 1, & k\neq 0, \end{cases}$$

where i, j, n_1, n_2 are in range of $[0, 7]$ so that $N_1 = N_2 = 8$, a_{n1}, a_{n2} are the constant coefficient. To collect the spectrum information along $2D$ images and patches, the $\tilde{X}_{pc}$ is flattened according to the frequency (F^2) from $1 \times 8 \times 8$ to the size of $64 \times 1 \times 1$, while the first number represents the channel and the last two refer to the length and the width, and every channel refers to the feature in a different frequency under the spectrum space. We first group the spectrum information from all image patches and generate the channel-wise DCT coefficient cube as $\tilde{X}_c \in \mathcal{R}^{64\times H/8\times W/8}$. Since the intensity value after DCT transformation would be mapped to a high range of scale in different frequency feature representations which is difficult for neural networks to handle and learn, we then implement another frequency-wise normalization (FN) channel by channel for every DCT coefficient cube and let them in the range of $[0, 1]$. Otherwise, the intensity values of the transformed data cube may range from zero to several million following the DCT transformation. To mitigate the impact of extreme outliers, we apply data clipping at the 5th and 95th percentiles before performing min-max normalization.

3.2. SIN-Seg with feature alignment

3.2.1. Segmentation framework

As shown in Fig. 2, we utilize *U-Net* as the backbone of our SIN-Seg framework. *U-Net* [2] is a widely used segmentation backbone that has shown convincing and robust performance on a large variety of medical image segmentation tasks. Here we adopt all default configurations used in the official implementations[1] with the input of 2D *RGB* images. Meanwhile, an extra encoder (i.e., the encoder of *U-Net*) is utilized to embed the DCT coefficient cube simultaneously. A shape alignment (the SA block in Fig. 2, combined with a dimension alignment by the Up block and a channel alignment by the CA block), is operated on the DCT coefficient cube before it goes through the encoder. In the same output scale of the *U-Net* encoder, the feature maps of the original image and DCT coefficient cube are concatenated as a fused feature map.

3.2.2. Feature alignment

We propose a new contrastive alignment module and conduct the feature alignment after the last down-sample of the *U-Net* encoder. Particularly, we first utilize an MLP layer to flatten the feature maps into a feature band with a size of 1×512. Denote the feature band in the frequency domain and spatial domain as $\tilde{F}$ and F, respectively. An alignment matrix (AM) can then be constructed as $F_{align} = F^{\top}\tilde{F} \in \mathcal{R}^{512\times 512}$. Inspired by CLIP [16], who proposed a novel Dual-Modality Learning, which forces their CLIP model to learn from two modalities: images and text. It employs two neural networks, one for processing images and another for processing text. The goal is to map these two different types of data into a shared embedding space where they can be directly compared. This function operates by pulling the embeddings of matching image-text pairs closer together in the shared space while pushing non-matching pairs apart. For instance, an image of a dog and its correct textual description "A dog playing in the park" are pulled closer, whereas mismatches like the same image with the text "A cat sleeping" are pushed apart. This means it can understand and categorize images it has never seen during training, based solely on its learned associations between text and images. For a broader explanation, such a contrastive training strategy could align the correlated features from different spaces to be in a shared new feature space, and those uncorrelated features to be pushed away in this new space. In our module, we assume that the corresponding features (i.e., $\tilde{F}_{:,:,i}$ and $F_{:,:,i}$) are more correlated, while the non-corresponding features (i.e., $\tilde{F}_{:,:,i}$ and $F_{:,:,j}$) are less correlated. In other words, the diagonal elements in F_{align}

[1] https://github.com/milesial/Pytorch-UNet.

should be dominated. So that all the non-corresponding features would be regarded as negative samples while those corresponding features are positive samples. To this end, a Binary Cross Entropy (BCE) loss is proposed to achieve this contrastive alignment process, and the loss function is as follows:

$$L_{Align}(\tilde{X},\ X^{RGB}) = BCE(F_{align},\ E), \tag{3}$$

where E is a diagonal matrix (DM) with a size of 512×512.

Loss function. The loss function for our proposed SIN-Seg framework consists of two parts, including the segmentation loss and the proposed feature alignment loss. Following previous methods [49,50], we use BCE loss and Dice loss together as the segmentation loss. Therefore, the whole loss function is formulated as:

$$L_{total} = L_{CE} + L_{Dice} + L_{align}, \tag{4}$$

4. Experiments

4.1. Datasets

We use four publically available datasets captured from different commonly used medical imaging devices, including the NeurIPS CellSeg 2022(CellSeg) dataset [20], the CHAOS-CT abdominal organ segmentation (CHAOS-CT) dataset [19], the medical segmentation decathlon heart (MSD-Heart) dataset [18], and a brain tumor segmentation dataset BraTS 2015 [17]in this study.

- **CellSeg**: The CellSeg dataset consists of 1000 microscope 2D image slices (i.e., 900 slices training and 100 slices testing) collected from 10 different organizations. It is a specialized dataset designed for advancing research in the field of cellular image analysis, aiding in understanding cellular structures and functions. It includes a wide range of images capturing various types of cells under different imaging conditions. All slices were manually labeled with 11 segmentation regions, such as yeast, adipocyte, brain cell, etc.
- **CHAOS-CT**: The CHAOS challenge (Combined (CT-MR) Healthy Abdominal Organ Segmentation) is a specialized collection of medical images designed for the evaluation and development of computer-aided diagnosis systems, particularly focusing on liver segmentation. We use the CT part of the challenge dataset, a series of abdominal CT scans, providing a comprehensive view of the liver and surrounding organs. These scans are sourced from different patients, offering a diverse range of liver shapes, sizes, and pathologies. It consists of 2875 CT slices from 40 different patients collected by the DEU hospital, where the liver regions were manually labeled by expert radiologists.
- **MSD-Heart**: The MSD-Heart dataset is part of the Medical Segmentation Decathlon (MSD), a comprehensive collection of datasets aimed at advancing the field of medical image segmentation. Specifically, the MSD-Heart dataset focuses on the segmentation of cardiac structures from MRI scans. This dataset includes a series of MRI scans that capture detailed images of the heart. These scans are sourced from a diverse patient population, encompassing a wide range of heart shapes, sizes, and pathologies. Such diversity is crucial for developing segmentation algorithms that are robust and effective across different patient demographics and clinical conditions. It consists of 2272 MRI slices from 30 subjects, where the experts manually labeled the left atrium.
- **Brain Tumor Segmentation**: The BraTS2015 (Brain Tumor Segmentation 2015) challenge dataset is a significant resource for brain tumor segmentation. It is a dataset for an annual competition that focuses on the segmentation of gliomas, a common type of brain tumor, from multimodal MRI scans. This dataset includes four different MRI modalities: T1, T1-contrast enhanced, T2, and FLAIR (Fluid Attenuated Inversion Recovery), providing a comprehensive view of the tumor and surrounding brain tissues. We used the T2 modality for experiments, including 35 3D MRI images. We generate 5000 2D image slices from these 3D MRI images for tumor segmentation, where 80% and 20% of image slices are utilized for framework training and validation, respectively.

In this study, the effects of subjects' age, gender, race, or any other variables on the results are not evaluated since the related information is not provided by the data provider. Details of the data description and preprocessing are shown below.

4.2. Implementation details

We first resize each image to a size of 128×128 through bilinear interpolation for network training, with training epochs as 300 and 75 epochs for early stop patience. In order to explore the lower bound contribution of the introduced spectral information, we refrain from using any data augmentation techniques in all of our experiments. We trained the module by using the Adam optimizer with a batch size of 20 and synchronized batch normalization. The initial learning rate was set to $1e^{-3}$ and decayed by $(1-\frac{current\_epoch}{max\_epoch})^{0.9}$ with an l_2 weight decay of $5e^{-4}$. All experiments were conducted based on PyTorch 1.7.1 and were deployed on a workstation with 2× NVIDIA TITAN RTX GPUs which owns 24 GB memory individually. It is worth mentioning that we didn't use all of the memory in two GPUs, the detailed occupied memory could be found in Fig. 5.

4.3. Baselines and evaluation metrics

We compared our proposed SIN-Seg framework with 8 influential U-Net and Transformer-based segmentation baselines, i.e., *U-Net* [2], UNet++ [23], ResUNet [21], AttUNet [22], UNeXt [24], MedT [28], MissFormer [29], FRCU-Net [38], XNet [42]. *U-Net* is a cutting-edge backbone framework for medical image segmentation, and UNet++, ResUNet, AttUNet, and UNeXt are four well-performing segmentation frameworks based on the *U-Net* backbone. TransUNet, MedT, and MissFormer are three models that take advantage of the ViT module, FRCU-Net and XNet are two prominent models that employ spectrum-based techniques to enhance medical image segmentation. We adopt two metrics to assess the performance of segmentation methods, including the Dice similarity coefficient (DSC, see as Eq. (5)), which are overlap-based metrics ranging from 0 to 1 and mean intersection over union (IoU, see as Eq. (6)), while X represents the set of pixels in the first segmentation (e.g., the algorithm's output), Y represents the set of pixels in the second segmentation (e.g., the ground truth). $|X \cap Y|$ is the cardinality of the intersection of sets X and Y (i.e., the number of pixels common to both segmentation). $|X| + |Y|$ are the cardinalities of sets X and Y, is the cardinality of the union of sets, $|X_i \cup Y_i|$ is the cardinality of the union of sets X_i and Y_i for the i_th class (i.e., the total number of pixels in both the predicted and ground truth segmentation for that class), respectively (i.e., the total number of pixels in each segmentation).

$$\mathrm{DSC} = \frac{2 \times |X \cap Y|}{|X| + |Y|}, \tag{5}$$

$$\mathrm{mIoU} = \frac{1}{N}\sum_{i=1}^{N}\frac{|X_i \cap Y_i|}{|X_i \cup Y_i|}, \tag{6}$$

4.4. Comparative experiments

Table 1 and Table 2 provides the performance of eight baseline methods and our SIN-Seg-based on UNet and Transformer respectively. It shows that our method outperforms all baselines substantially in terms of both metrics of the dataset. Compared to the results based on *U-Net* and *TransUNet* two foundation models, our proposed fusion framework could enhance the baseline models and achieve superior segmentation

Table 1
Quantitative results of different methods on CellSef and MSD-Heart datasets. The best results are shown in bolded font and the second best results are underlined. The values of DSC and IoU are in percentage terms.

Frameworks	CellSeg				MSD-Heart			
	DSC↑	IoU↑	HD95↓	ASD↓	DSC↑	IoU↑	HD95↓	ASD↓
FRCU-Net[a]	$86.17_{\pm 0.23}$	$74.33_{\pm 0.16}$	$2.23_{\pm 0.26}$	$1.30_{\pm 0.65}$	$91.09_{\pm 1.12}$	$85.13_{\pm 0.16}$	$5.80_{\pm 0.15}$	$2.93_{\pm 0.45}$
XNet[b]	$85.63_{\pm 0.56}$	$77.48_{\pm 1.08}$	$3.17_{\pm 0.35}$	$1.92_{\pm 0.54}$	$89.10_{\pm 0.46}$	$84.09_{\pm 1.04}$	$4.78_{\pm 0.74}$	$\mathbf{2.02_{\pm 0.92}}$
U-Net	$85.56_{\pm 0.54}$	$71.23_{\pm 0.39}$	$4.15_{\pm 1.4}$	$2.08_{\pm 0.49}$	$90.63_{\pm 1.14}$	$83.55_{\pm 0.64}$	$3.76_{\pm 0.40}$	$2.11_{\pm 0.66}$
UNet++	$83.90_{\pm 1.02}$	$70.07_{\pm 0.85}$	$4.46_{\pm 0.54}$	$2.20_{\pm 0.68}$	$91.55_{\pm 0.85}$	$85.29_{\pm 0.94}$	$6.71_{\pm 0.54}$	$5.36_{\pm 0.36}$
ResUNet	$84.08_{\pm 0.71}$	$71.29_{\pm 0.08}$	$9.80_{\pm 0.40}$	$4.41_{\pm 0.52}$	$87.51_{\pm 0.42}$	$79.30_{\pm 0.54}$	$7.62_{\pm 0.13}$	$5.90_{\pm 0.24}$
AttUNet	$82.50_{\pm 0.33}$	$\mathbf{78.25_{\pm 1.69}}$	$6.71_{\pm 0.58}$	$4.34_{\pm 0.29}$	$88.78_{\pm 0.44}$	$83.30_{\pm 0.40}$	$6.31_{\pm 0.24}$	$2.16_{\pm 0.64}$
UNeXt	$84.48_{\pm 0.49}$	$72.08_{\pm 0.14}$	$3.19_{\pm 0.33}$	$1.49_{\pm 0.54}$	$74.71_{\pm 1.04}$	$73.60_{\pm 0.87}$	$5.09_{\pm 0.39}$	$3.72_{\pm 0.19}$
UNet+SINSeg	$\mathbf{86.55_{\pm 0.60}}$	$73.16_{\pm 0.69}$	$\mathbf{1.67_{\pm 0.50}}$	$\mathbf{1.15_{\pm 0.19}}$	$\mathbf{92.50_{\pm 0.91}}$	$\mathbf{88.61_{\pm 1.04}}$	$\mathbf{2.12_{\pm 0.60}}$	$2.63_{\pm 0.34}$
TransUNet	$86.92_{\pm 1.04}$	$74.89_{\pm 0.80}$	$5.16_{\pm 0.29}$	$2.39_{\pm 0.70}$	$73.86_{\pm 0.44}$	$69.53_{\pm 0.20}$	$6.10_{\pm 0.66}$	$4.23_{\pm 0.19}$
MedT	$84.91_{\pm 0.60}$	$76.59_{\pm 0.34}$	$3.12_{\pm 0.86}$	$2.02_{\pm 0.16}$	$\mathbf{81.51_{\pm 1.06}}$	$\mathbf{76.32_{\pm 1.09}}$	$2.38_{\pm 0.56}$	$2.45_{\pm 0.33}$
MissFormer	$83.14_{\pm 1.68}$	$75.08_{\pm 0.50}$	$3.11_{\pm 0.90}$	$1.98_{\pm 0.26}$	$77.52_{\pm 0.60}$	$75.55_{\pm 0.20}$	$6.62_{\pm 0.80}$	$4.36_{\pm 0.55}$
TransUNet+SINSeg	$\mathbf{89.23_{\pm 1.12}}$	$\mathbf{81.08_{\pm 0.26}}$	$\mathbf{2.09_{\pm 0.12}}$	$\mathbf{1.19_{\pm 0.20}}$	$78.70_{\pm 0.58}$	$75.29_{\pm 0.16}$	$\mathbf{2.15_{\pm 0.20}}$	$\mathbf{2.29_{\pm 0.33}}$

[a] Indicates that we use only the architecture of the method without incorporating any extra pre-trained models to ensure fairness.
[b] Stands for we only use the fully-supervised setting in our comparison.

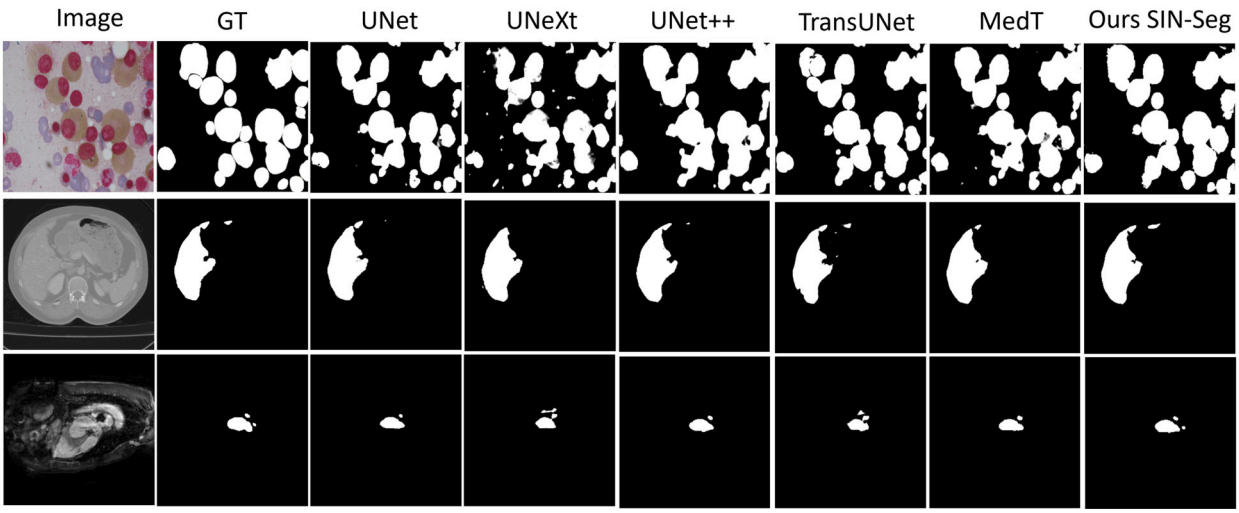


Fig. 3. Visualization of the segmentation results produced by our frameworks and typical baselines on the CellSeg (row 1), CHAOS-CT (row 2), and MSD-Heart (row 3) datasets. * For "Ours SIN-Seg", it is the predicted results with the UNet as the backbone.

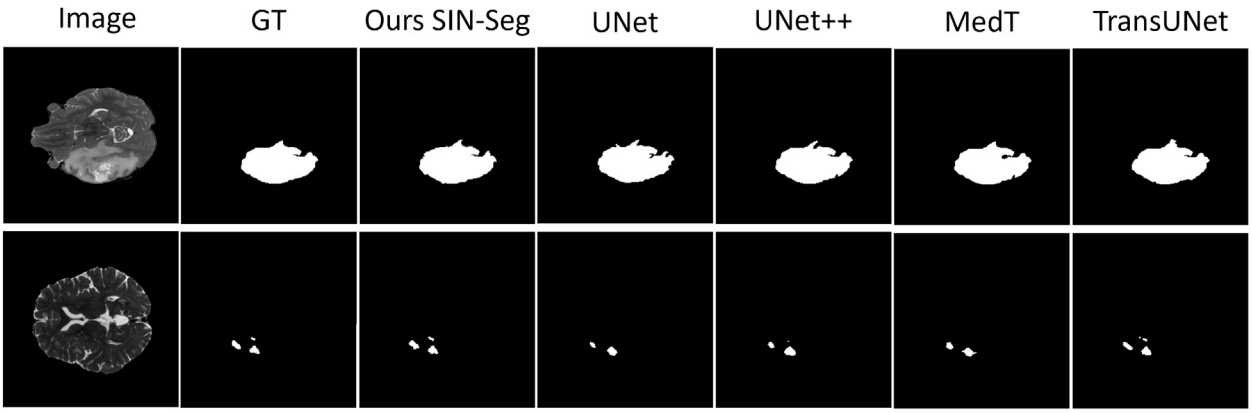
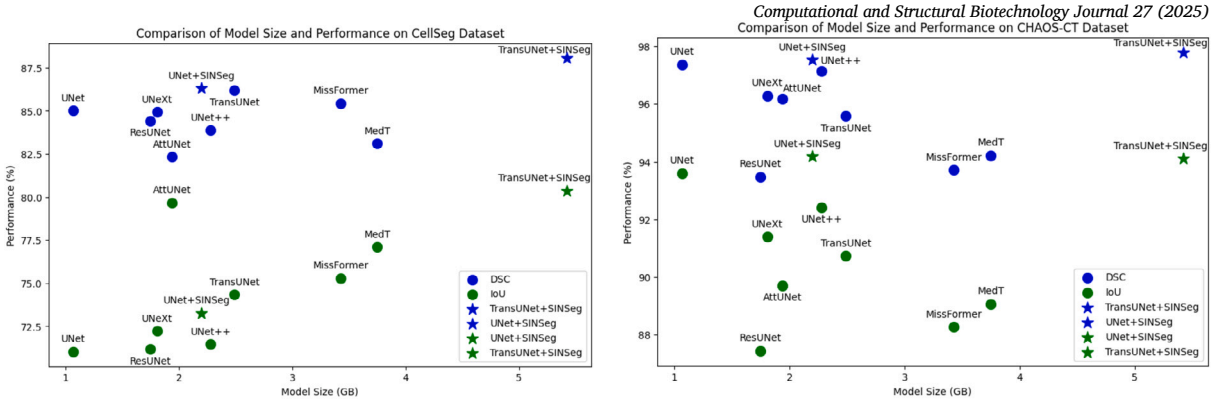


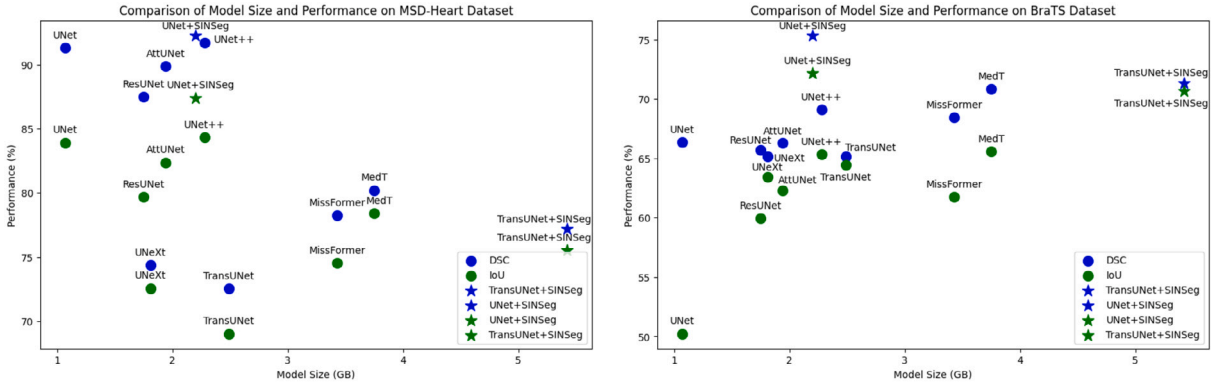
Fig. 4. Visualization of the representative segmentation results produced by our frameworks and typical baselines on the BraTS dataset. The first column represents when the lesion is large and the second column illustrates the situation when the tumor is small and discrete distributed. * For "Ours SIN-Seg", it is the predicted results with the UNet as the backbone.

results, which shows the importance of introducing spectrum information as a complement to spatial information in deep neural networks for segmentation tasks. We also visualized the segmentation results among four datasets in Fig. 3 and Fig. 4, our visualization results in two figures are both designed based on UNet. For the first three datasets, the visualization results illustrate that the results produced by our SIN-Seg



(a) Performance and Model Size for CellSeg dataset

(b) Performance and Model Size for CHAOS-CT dataset



(c) Performance and Model Size for MSD-Heart dataset

(d) Performance and Model Size for BraTS dataset

Fig. 5. Illustration of the correlation between the model complexity and the performance.

Table 2
Quantitative results of different methods on CHAOS-CT and BraTS datasets. The best results are shown in bolded font and the second best results are underlined. The values of DSC and IoU are in percentage terms.

Frameworks	CHAOS-CT				BraTS			
	DSC↑	IoU↑	HD95↓	ASD↓	DSC↑	IoU↑	HD95↓	ASD↓
FRCU-Net[a]	$97.21_{\pm 0.81}$	$93.66_{\pm 0.72}$	$3.13_{\pm 0.11}$	$2.71_{\pm 0.39}$	$69.90_{\pm 0.46}$	$58.29_{\pm 0.59}$	$5.63_{\pm 0.91}$	$4.22_{\pm 0.52}$
XNet[b]	$95.92_{\pm 1.01}$	$\mathbf{94.13_{\pm 0.57}}$	$3.29_{\pm 0.31}$	$\mathbf{2.18_{\pm 0.50}}$	$72.11_{\pm 0.35}$	$66.30_{\pm 0.64}$	$7.11_{\pm 0.11}$	$5.91_{\pm 0.29}$
U-Net	$97.70_{\pm 0.65}$	$93.87_{\pm 0.91}$	$3.69_{\pm 0.32}$	$2.40_{\pm 0.40}$	$68.25_{\pm 0.63}$	$52.22_{\pm 1.90}$	$7.73_{\pm 0.47}$	$5.72_{\pm 0.90}$
UNet++	$96.30_{\pm 0.61}$	$93.55_{\pm 1.01}$	$4.66_{\pm 0.60}$	$4.10_{\pm 0.35}$	$98.91_{\pm 0.62}$	$64.50_{\pm 0.71}$	$7.61_{\pm 0.58}$	$6.82_{\pm 0.26}$
ResUNet	$94.28_{\pm 0.56}$	$88.26_{\pm 0.39}$	$5.90_{\pm 0.11}$	$4.65_{\pm 0.52}$	$65.32_{\pm 0.47}$	$60.58_{\pm 0.40}$	$8.82_{\pm 0.88}$	$8.03_{\pm 0.45}$
AttUNet	$95.77_{\pm 0.82}$	$89.35_{\pm 1.51}$	$5.12_{\pm 0.31}$	$4.84_{\pm 0.66}$	$66.92_{\pm 0.31}$	$63.82_{\pm 0.78}$	$8.22_{\pm 0.63}$	$6.38_{\pm 0.96}$
UNeXt	$95.02_{\pm 0.35}$	$91.52_{\pm 0.31}$	$7.79_{\pm 0.22}$	$6.18_{\pm 0.53}$	$65.80_{\pm 0.37}$	$65.19_{\pm 1.51}$	$10.52_{\pm 1.36}$	$10.05_{\pm 0.81}$
UNet+SINSeg	$\mathbf{98.15_{\pm 0.49}}$	$\mathbf{95.20_{\pm 0.57}}$	$\mathbf{2.66_{\pm 0.55}}$	$2.44_{\pm 0.46}$	$\mathbf{75.63_{\pm 0.30}}$	$\mathbf{73.14_{\pm 0.71}}$	$\mathbf{4.34_{\pm 0.59}}$	$\mathbf{3.62_{\pm 0.46}}$
TransUNet	$95.10_{\pm 0.31}$	$91.44_{\pm 0.62}$	$\mathbf{3.35_{\pm 0.13}}$	$\mathbf{3.06_{\pm 0.71}}$	$66.28_{\pm 0.36}$	$65.37_{\pm 0.45}$	$5.88_{\pm 0.62}$	$4.90_{\pm 0.39}$
MedT	$94.90_{\pm 1.32}$	$90.32_{\pm 0.41}$	$4.89_{\pm 0.85}$	$4.62_{\pm 0.23}$	$69.45_{\pm 0.39}$	$64.29_{\pm 0.30}$	$6.80_{\pm 0.21}$	$6.14_{\pm 0.63}$
MissFormer	$93.06_{\pm 0.51}$	$86.54_{\pm 0.31}$	$5.61_{\pm 0.29}$	$4.66_{\pm 0.40}$	$69.16_{\pm 0.81}$	$61.98_{\pm 0.47}$	$7.29_{\pm 0.51}$	$6.65_{\pm 0.83}$
TransUNet+SINSeg	$\mathbf{96.38_{\pm 0.27}}$	$\mathbf{93.41_{\pm 0.56}}$	$4.26_{\pm 0.75}$	$3.80_{\pm 0.47}$	$\mathbf{70.85_{\pm 1.01}}$	$\mathbf{72.06_{\pm 0.55}}$	$\mathbf{5.05_{\pm 0.33}}$	$\mathbf{4.18_{\pm 0.72}}$

[a] Represents that we use only the architecture of the method without incorporating any extra pre-trained models to ensure fairness.
[b] Stands for we only use the fully-supervised setting in our comparison.

framework are more similar to the ground truths than those generated by other typical baselines, especially for some of the detailed edges. According to Fig. 4, we demonstrated two situations from huge tumors and small with discretely distributed tumors. Our proposed SIN-Seg could wisely handle two difficult situations in one target subject, which represents our novel framework can learn intrinsic semantic well.

4.5. Ablation study

We conducted an ablation study on four datasets to evaluate the necessity and importance of each component in our framework. Table 3 shows that our SIN-Seg framework improves the DSC and IoU substantially compared with U-Net by just using pure spatial or spectral information for the dataset, which is due to insufficient information. Therefore, both spatial and spectral information play important roles in medical image segmentation. However, a naive combination of the information from different spaces is also unreasonable. One simple U-Net model cannot handle two types of information space. Alignment of the features and mapping them into a shared space for synchronization is crucial, otherwise, the performance would even be worse. The comparison between SINSeg and SINSeg without feature alignment indicates the contributions provided by the proposed feature alignment loss.

Table 3
Ablation studies of our proposed SIN-Seg framework on the other three datasets. The best results are shown in bolded font.

Settings	CellSeg		CHAOS-CT		MSD-Heart		BraTS	
	DSC	IoU	DSC	IoU	DSC	IoU	DSC	IoU
U-Net+Pure Spatial	$85.56_{\pm 0.54}$	$71.23_{\pm 0.39}$	$97.70_{\pm 0.65}$	$93.87_{\pm 0.91}$	$90.63_{\pm 1.14}$	$83.55_{\pm 0.64}$	$68.25_{\pm 0.63}$	$52.22_{\pm 1.90}$
U-Net+Pure Spectrum	$71.52_{\pm 0.89}$	$57.47_{\pm 1.92}$	$95.26_{\pm 0.52}$	$92.46_{\pm 0.37}$	$87.58_{\pm 0.88}$	$82.76_{\pm 0.58}$	$65.98_{\pm 1.31}$	$52.47_{\pm 0.29}$
U-Net+Joint wo Alignment	$78.87_{\pm 0.60}$	$60.09_{\pm 1.39}$	$96.35_{\pm 0.49}$	$90.06_{\pm 1.61}$	$89.02_{\pm 1.31}$	$82.54_{\pm 0.86}$	$50.81_{\pm 1.84}$	$47.31_{\pm 2.15}$
U-Net+SINSeg	$\mathbf{86.55_{\pm 0.60}}$	$\mathbf{73.16_{\pm 0.69}}$	$\mathbf{98.15_{\pm 0.49}}$	$\mathbf{95.20_{\pm 0.57}}$	$\mathbf{92.50_{\pm 0.91}}$	$\mathbf{88.61_{\pm 1.04}}$	$\mathbf{75.63_{\pm 0.30}}$	$\mathbf{73.14_{\pm 0.71}}$
TransUNet+Pure Spatial	$86.92_{\pm 1.04}$	$74.89_{\pm 0.80}$	$95.10_{\pm 0.31}$	$91.44_{\pm 0.62}$	$73.86_{\pm 0.44}$	$69.53_{\pm 0.20}$	$66.28_{\pm 0.36}$	$65.37_{\pm 0.45}$
TransUNet+Pure Spectrum	$73.88_{\pm 1.03}$	$52.31_{\pm 0.89}$	$89.63_{\pm 0.67}$	$82.03_{\pm 0.81}$	$73.09_{\pm 0.62}$	$61.30_{\pm 1.42}$	$54.23_{\pm 1.21}$	$53.27_{\pm 1.29}$
TransUNet+Joint wo Alignment	$75.04_{\pm 0.18}$	$65.42_{\pm 0.59}$	$92.37_{\pm 1.29}$	$82.19_{\pm 0.88}$	$73.29_{\pm 1.24}$	$65.36_{\pm 1.04}$	$62.39_{\pm 0.78}$	$59.19_{\pm 0.55}$
TransUNet+SINSeg	$\mathbf{89.23_{\pm 1.12}}$	$\mathbf{81.08_{\pm 0.26}}$	$\mathbf{96.38_{\pm 0.27}}$	$\mathbf{93.41_{\pm 0.56}}$	$\mathbf{78.70_{\pm 0.58}}$	$\mathbf{75.29_{\pm 0.16}}$	$\mathbf{70.85_{\pm 1.01}}$	$\mathbf{72.06_{\pm 0.55}}$

Table 4
Extra experiments to concern the potential overfitting problem for proposed SINSeg.

Datasets	Train		Validation	
	DSC	IoU	DSC	IoU
CellSeg	$91.06_{\pm 0.53}$	$77.18_{\pm 0.62}$	$\mathbf{86.55_{\pm 0.60}}$	$\mathbf{73.16_{\pm 0.69}}$
CHAOS-CT	$97.42_{\pm 0.68}$	$96.18_{\pm 0.38}$	$\mathbf{98.15_{\pm 0.49}}$	$\mathbf{95.20_{\pm 0.57}}$
MSD-Heart	$94.97_{\pm 0.39}$	$89.02_{\pm 0.58}$	$\mathbf{92.50_{\pm 0.91}}$	$\mathbf{88.61_{\pm 1.04}}$
BraTS	$81.26_{\pm 0.57}$	$75.69_{\pm 0.41}$	$\mathbf{75.63_{\pm 0.30}}$	$\mathbf{73.14_{\pm 0.71}}$

5. Discussion and limitation

5.1. Discussion

Since our proposed method employs deep convolutional neural networks (DCNNs), we conducted additional experiments to discuss some of the main issues considered with DCNNs. We first explore the potential overfitting issue by comparing the performance between the training step and the validation step. According to the Table 4, even though the validation performances are all lower than on the training step, the gap is slight which could demonstrate that our proposed SINSeg does not have an overfitting problem. Meanwhile, the proposed module naturally introduces more trainable parameters. In order to illustrate the effectiveness, we plot the correlation between model size and the performance in Fig. 5. It clearly shows the superiority of our proposed method, when employing the SINSeg module in the original UNet model, it could always obtain the best performance in a small size of the occupied GPU memory. Even though the TransUNet+SINSeg is the biggest model of all, it is not always the best. We conjecture that it is because of the limitation of the dataset size, that larger models always need a richer dataset to handle the extra trainable parameters.

5.2. Limitation

Due to the reliance on commonly used medical datasets and the high cost associated with acquiring larger datasets, assessing the scalability of our proposed method remains a challenge. Additionally, our study is limited to the medical image segmentation task. Future work could explore other critical tasks within the medical domain, such as classification, regression, and reconstruction, to fully evaluate the applicability and effectiveness of our approach.

6. Conclusion

In this paper, we propose a spectrum information-based feature-enhanced (SIN) model that combines spectrum and spatial information for different segmentation tasks. Experimental results demonstrate the effectiveness and superiority of our proposed model. According to our comprehensive analysis, spectral information plays an important role in medical image segmentation tasks and should be fully considered. The semantic features yielded from the spectrum space should be aligned because feature variances, resulting from the inconsistent frequency-related settings of medical imaging modalities, exist on the segmentation ROIs. In fact, we introduce more parameters compared to the other models, but most of the feature channels in the spectrum space may be surplus [35]. In the future, we will plan to use an advanced feature selection mechanism for spatial and spectrum feature spaces.

CRediT authorship contribution statement

Siyuan Dai: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Conceptualization. **Kai Ye:** Writing – review & editing, Visualization, Investigation, Formal analysis, Conceptualization. **Charlie Zhan:** Writing – review & editing, Conceptualization. **Haoteng Tang:** Writing – review & editing, Supervision, Investigation, Funding acquisition, Conceptualization. **Liang Zhan:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This study was partially supported by the National Science Foundation (Grant No. IIS 2045848), the Presidential Research Fellowship in the Department of Computer Science at the University of Texas Rio Grande Valley. In the revision process of this research paper, the authors received assistance from ChatGPT, an AI language model developed by OpenAI. ChatGPT provided valuable insights and suggestions, which contributed to the refinement and improvement of the manuscript. We acknowledge ChatGPT's assistance in enhancing the clarity and coherence of the paper's content.

References

[1] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2015. p. 3431–40.
[2] Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference. Proceedings part III, vol. 18. Springer; 2015. p. 234–41.
[3] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. p. 770–8.
[4] Fu X, Sun Z, Tang H, Zou EM, Huang H, Wang Y, et al. 3d bi-directional transformer u-net for medical image segmentation. Front Big Data 2023;5:1080715.
[5] Jia H, Tang H, Ma G, Cai W, Huang H, Zhan L, et al. A convolutional neural network with pixel-wise sparse graph reasoning for covid-19 lesion segmentation in ct images. Comput Biol Med 2023:106698.

[6] Ye K, Tang H, Dai S, Guo L, Liu JY, Wang Y, et al. Bidirectional mapping with contrastive learning on multimodal neuroimaging data. In: International conference on medical image computing and computer-assisted intervention. Springer; 2023. p. 138–48.
[7] Guo Z, Li X, Huang H, Guo N, Li Q. Deep learning-based image segmentation on multimodal medical imaging. IEEE Trans Radiat Plasma Med Sci 2019;3(2):162–9.
[8] Wang W, Chen C, Ding M, Yu H, Zha S, Li J. Transbts: multimodal brain tumor segmentation using transformer. In: Medical image computing and computer assisted intervention–MICCAI 2021: 24th international conference. Proceedings, part I, vol. 24. Springer; 2021. p. 109–19.
[9] Zhong Y, Li B, Tang L, Kuang S, Wu S, Ding S. Detecting camouflaged object in frequency domain. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2022. p. 4504–13.
[10] Xu K, Qin M, Sun F, Wang Y, Chen Y-K, Ren F. Learning in the frequency domain. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2020. p. 1740–9.
[11] Kruse FA, Lefkoff A, Boardman J, Heidebrecht K, Shapiro A, Barloon P, et al. The spectral image processing system (sips)—interactive visualization and analysis of imaging spectrometer data. Remote Sens Environ 1993;44(2–3):145–63.
[12] Xu P, Zhu X, Clifton DA. Multimodal learning with transformers: a survey. IEEE Trans Pattern Anal Mach Intell 2023;45(10):12113–32.
[13] Zhao WX, Zhou K, Li J, Tang T, Wang X, Hou Y, et al. A survey of large language models. arXiv preprint. arXiv:2303.18223, 2023.
[14] Wang Y, Chen X, Cao L, Huang W, Sun F, Wang Y. Multimodal token fusion for vision transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2022. p. 12186–95.
[15] Lee J, Dabagia M, Dyer E, Rozell C. Hierarchical optimal transport for multimodal distribution alignment. Adv Neural Inf Process Syst 2019;32.
[16] Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning transferable visual models from natural language supervision. In: International conference on machine learning; 2021. p. 8748–63.
[17] Menze BH, Jakab A, Bauer S, Kalpathy-Cramer J, Farahani K, Kirby J, et al. The multimodal brain tumor image segmentation benchmark (brats). IEEE Trans Med Imaging 2014;34(10):1993–2024.
[18] Antonelli M, Reinke A, Bakas S, Farahani K, Kopp-Schneider A, Landman BA, et al. The medical segmentation decathlon. Nat Commun 2022;13(1):4128.
[19] Kavur AE, Gezer NS, Barış M, Aslan S, Conze P-H, Groza V, et al. CHAOS challenge - combined (CT-MR) healthy abdominal organ segmentation. Med Image Anal 2021;69:101950. https://doi.org/10.1016/j.media.2020.101950. http://www.sciencedirect.com/science/article/pii/S1361841520303145.
[20] Ma J, Xie R, Ayyadhury S, Ge C, Gupta A, Gupta R, et al. The multi-modality cell segmentation challenge: towards universal solutions. arXiv:2308.05864, 2023.
[21] Zhang Z, Liu Q, Wang Y. Road extraction by deep residual u-net. IEEE Geosci Remote Sens Lett 2018;15(5):749–53.
[22] Oktay O, Schlemper J, Le Folgoc L, Lee M, Heinrich M, Misawa K, et al. Attention u-net: learning where to look for the pancreas. In: Medical imaging with deep learning; 2022.
[23] Zhou Z, Rahman Siddiquee MM, Tajbakhsh N, Liang J. Unet++: a nested u-net architecture for medical image segmentation. In: Deep learning in medical image analysis and multimodal learning for clinical decision support: 4th international workshop, DLMIA 2018, and 8th international workshop, ML-CDS 2018, held in conjunction with MICCAI 2018. Proceedings, vol. 4. Granada, Spain: Springer; 2018. p. 3–11.
[24] Valanarasu JMJ, Patel VM. Unext: mlp-based rapid medical image segmentation network. In: International conference on medical image computing and computer-assisted intervention. Springer; 2022. p. 23–33.
[25] Chen J, Lu Y, Yu Q, Luo X, Adeli E, Wang Y, et al. Transunet: transformers make strong encoders for medical image segmentation. arXiv preprint. arXiv:2102.04306, 2021.
[26] Han K, Wang Y, Chen H, Chen X, Guo J, Liu Z, et al. A survey on vision transformer. IEEE Trans Pattern Anal Mach Intell 2022;45(1):87–110.
[27] Khan S, Naseer M, Hayat M, Zamir SW, Khan FS, Shah M. Transformers in vision: a survey. ACM Comput Surv 2022;54(10s):1–41.
[28] Valanarasu JMJ, Oza P, Hacihaliloglu I, Patel VM. Medical transformer: gated axial-attention for medical image segmentation. In: Medical image computing and computer assisted intervention–MICCAI 2021: 24th international conference. Proceedings part I, vol. 24. Springer; 2021. p. 36–46.
[29] Huang X, Deng Z, Li D, Yuan X, Fu Y. Missformer: an effective transformer for 2d medical image segmentation. IEEE Trans Med Imaging 2022;42(5):1484–94.
[30] Cuthill I. Camouflage. J Zool 2019;308(2):75–92.
[31] Stevens M, Merilaita S. Animal camouflage: current issues and new perspectives. Philos Trans R Soc Lond B, Biol Sci 2009;364(1516):423–7.
[32] Gueguen L, Sergeev A, Kadlec B, Liu R, Yosinski J. Faster neural networks straight from jpeg. Adv Neural Inf Process Syst 2018;31.
[33] Ehrlich M, Davis LS. Deep residual learning in the jpeg transform domain. In: Proceedings of the IEEE/CVF international conference on computer vision; 2019. p. 3484–93.
[34] Chen W, Wilson J, Tyree S, Weinberger KQ, Chen Y. Compressing convolutional neural networks in the frequency domain. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining; 2016. p. 1475–84.
[35] Liu Z, Xu J, Peng X, Xiong R. Frequency-domain dynamic pruning for convolutional neural networks. Adv Neural Inf Process Syst 2018;31.
[36] Wang Y, Xu C, You S, Tao D, Xu C. Cnnpack: packing convolutional neural networks in the frequency domain. Adv Neural Inf Process Syst 2016;29.
[37] Qin Z, Zhang P, Wu F, Li X. Fcanet: frequency channel attention networks. In: Proceedings of the IEEE/CVF international conference on computer vision; 2021. p. 783–92.
[38] Azad R, Bozorgpour A, Asadi-Aghbolaghi M, Merhof D, Escalera S. Deep frequency re-calibration u-net for medical image segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision; 2021. p. 3274–83.
[39] Huang J, Guan D, Xiao A, Lu S. Fsdr: frequency space domain randomization for domain generalization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2021. p. 6891–902.
[40] Liu P, Zhang H, Zhang K, Lin L, Zuo W. Multi-level wavelet-cnn for image restoration. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops; 2018. p. 773–82.
[41] Zheng B, Yuan S, Yan C, Tian X, Zhang J, Sun Y, et al. Learning frequency domain priors for image demoireing. IEEE Trans Pattern Anal Mach Intell 2021;44(11):7705–17.
[42] Zhou Y, Huang J, Wang C, Song L, Yang G. Xnet: wavelet-based low and high frequency fusion networks for fully- and semi-supervised semantic segmentation of biomedical images. In: Proceedings of the IEEE/CVF international conference on computer vision; 2023. p. 21085–96.
[43] Zhang Y, Sidibé D, Morel O, Mériaudeau F. Deep multimodal fusion for semantic image segmentation: a survey. Image Vis Comput 2021;105:104042.
[44] Lahat D, Adali T, Jutten C. Multimodal data fusion: an overview of methods, challenges, and prospects. Proc IEEE 2015;103(9):1449–77.
[45] Huang K, Shi B, Li X, Li X, Huang S, Li Y. Multi-modal sensor fusion for auto driving perception: a survey. arXiv preprint. arXiv:2202.02703, 2022.
[46] Tan W, Tiwari P, Pandey HM, Moreira C, Jaiswal AK. Multimodal medical image fusion algorithm in the era of big data. Neural Comput Appl 2020:1–21.
[47] Yuan X, Lin Z, Kuen J, Zhang J, Wang Y, Maire M, et al. Multimodal contrastive training for visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2021. p. 6995–7004.
[48] Mustafa B, Riquelme C, Puigcerver J, Jenatton R, Houlsby N. Multimodal contrastive learning with limoe: the language-image mixture of experts. Adv Neural Inf Process Syst 2022;35:9564–76.
[49] Wei J, Hu Y, Li G, Cui S, Kevin Zhou S, Li Z. Boxpolyp: boost generalized polyp segmentation using extra coarse bounding box annotations. In: Medical image computing and computer assisted intervention–MICCAI 2022: 25th international conference. Proceedings, part III. Springer; 2022. p. 67–77.
[50] Fan D-P, Ji G-P, Zhou T, Chen G, Fu H, Shen J, et al. Pranet: parallel reverse attention network for polyp segmentation. In: Medical image computing and computer assisted intervention–MICCAI 2020: 23rd international conference. Proceedings, part VI, vol. 23. Springer; 2020. p. 263–73.